Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-03 17:04:03
REFERENCE NO: 208

Author Names & Affiliations

- Ramunas Stepanauskas - Bigelow Laboratory for Ocean Sciences

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

biology/environmental microbiology

Title of Submission

Reference databases for microbiome studies

Abstract (maximum ~200 words).

Microbiome studies increasingly rely on Big Data approaches applied on massive “omics” data sets (genomics, transcriptomics, proteomics, metabololics, etc.). This involves three key components of cyber infrastructure: computational hardware, bioinformatics software, and reference databases. While the quality and accessibility of the first two components have been steadily improving, reference databases are increasingly becoming the main bottleneck in the accurate and efficient use of microbiome data. Relevant reference databases are essential in the computational processing of omics data. To a large extent, they determine the ultimate outcomes of microbiome studies.
There are at least two main reasons behind the omics reference database challenges, each of which call for different solutions.
1. Existing, public databases and portals for omics data, such as GenBank and IMG (DOE JGI) were set up over a decade ago and do not scale to the size of omics studies of today.
2. Challenges in quality control and quality tracking of reference data, such as genome assemblies and gene annotations, which limits the value of omics data and may compromise the accuracy of omics studies.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Microbiome studies increasingly rely on Big Data approaches applied on massive “omics” data sets (genomics, transcriptomics, proteomics, metabololics, etc.). This involves three key components of cyber infrastructure: computational hardware, bioinformatics software, and reference databases. While the quality and accessibility of the first two components have been steadily improving, reference databases are increasingly becoming the main bottleneck in the accurate and efficient use of microbiome data. Relevant reference databases are essential in the computational processing of omics data. To a large extent, they determine the ultimate outcomes of microbiome studies.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

There are at least two main reasons behind the omics reference database challenges, each of which call for different solutions.
1. Existing, public databases and portals for omics data, such as GenBank and IMG (DOE JGI) were set up over a decade ago and do not scale to the size of omics studies of today. For some larger projects, data submission to these portals has become virtually impossible, due to insufficient throughput, forcing scientists to establish their own data portals. Data submission to public databases is also impeded by the limited flexibility of data and metadata formats. This causes fragmentation, limited accessibility and diminished value of unique and expensive reference data sets. At this point it remains unclear whether the situation is fixable by the old, centralized data storage approach. A more viable approach may be the development of data search algorithms that centralize data access while abandoning the ambition of centralizing data storage.
2. Challenges in quality control and quality tracking of reference data, such as genome assemblies and gene annotations, which limits the value of omics data and may compromise the accuracy of omics studies. In the case of de novo genome assemblies, benchmarking of entire workflows for the frequency of common artifacts (e.g. misassemblies, chimeras, contamination) and their documentation should be a routine. This is especially urgent in the case of metagenome assemblies and binning, which are generating large datasets with limited quality control and may compromise the interpretation of diverse omics results in subsequent studies. In the case of gene annotations, the insufficient effort placed into experimental verification of gene functional assignments has been discussed for many years and still remains to be addressed.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

Failures of current and prior (e.g. CAMERA) initiatives to address the reference database challenges should be analyzed, in order to better understand how ther management of future initiatives could be improved. Particular useful may be searches for new ways to maintain a constant dialogue between database developers, end users and funding sources, with a more apparent accountability of database developers to the stakeholders.

Consent Statement